Can Organic Chemistry Literature Enable Machine Learning Yield Prediction ?

J. Schleinitz*,[1, 2, a)] M. Langevin*,[2, 3, b)] Y. Smail,[4] B. Wehnert,[4] L. Grimaud,[1, c)] and R. Vuilleumier[2, d)]

[1)] *LBM, Département de chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005, Paris, France*

[2)] *PASTEUR, Département de chimie, École Normale Supérieure, PSL University, Sorbonne Université, CNRS, 75005, Paris, France*

[3)] *Molecular Design Sciences - Integrated Drug Discovery, Sanofi R%D, 94400, Vitry-Sur-Seine, France*

[4)] *UPMC, PSL University, Sorbonne Université, CNRS, 75005, Paris, France*

(Dated: 23 March 2022)

Synthetic yield prediction using machine learning is intensively studied. While previous work focused on an ideal use case, High-Throughput Experiment datasets, predicting yields using literature data remains elusive. We built a large literature-based dataset of more than a thousand reactions, focusing on the activation of carbon-oxygen bonds of phenol derivatives under nickel catalysis. Detailed reaction conditions and associated yields were manually curated and stored in an open-access database. We assessed the performances of state-of-the-art machine learning models on this dataset, and explored their ability to realize predictions on novel publications, coupling partners and substrates. Our work shows that on well-designed yield prediction tasks, machine learning can have practical applications, and provides a unique public database for further improvements of these methods adapted to literature chemical data.

Keywords: Machine learning - Dataset - Reaction yield prediction

[a)] * Those authors contributed equally to this work; Electronic mail: jules.schleinitz@ens.psl.eu

[b)] * Those authors contributed equally to this work; Electronic mail: maxime.langevin@sanofi.com

[c)] Electronic mail: laurence.grimaud@ens.psl.edu

[d)] Electronic mail: rodolphe.vuilleumier@ens.psl.eu

Machine learning (ML) algorithms learn complex functions from data. As it can leverage existing data to perform *in silico* approximations of costly experimental processes, ML applications have sparked strong interest in chemical sciences. While ML has already made a significant impact in drug development[1,2], synthetizability assessment of small molecules[3] or Computer Aided Synthesis Planning[4], the ability of ML to predict a reaction yield from its experimental conditions remains a major challenge[5] that is intensively studied[6,7]. Advances on reaction yield prediction would have a major impact on organic synthesis by significantly reducing cost, time and resources necessary to synthesize novel chemicals.

Progress in ML is markedly driven by the increasing access to data. Thus, currently available datasets shape the evolution of ML for reaction yield prediction. Despite this, there are very few publicly available and easily operable datasets of chemical reactions with associated yields (Table S1). One of those few public datasets is the United State Patent and Trademark Office (USPTO) dataset[8] that covers a wide range of chemical reactions extracted from patents. USPTO data is extremely diverse and suffers from a selection bias as only successful reactions tend to be reported in patents. ML has shown poor performance predicting yields on this dataset ($R^2 < 0.2$[7]). In addition, two High Throughput Experiment (HTE) datasets, one of a Suzuki-Miyaura coupling,[9] and one of a palladium-catalyzed Buchwald-Hartwig cross-coupling,[10] are available in the literature. State-of-the-art modeling performs extremely well on those high-quality datasets ($R^2 > 0.8$[7,10]), but the extremely focused chemical reaction space covered by HTE limits the predictions to a narrow scope of experimental conditions and reactants.

While those datasets have enabled rapid progress of ML for yield prediction, there is a need of publicly available datasets[11,12] more representative of chemical reaction data available in the literature or used by chemists in their everyday work. To address this, we built a literature-mined reaction dataset that focuses on a specific reaction class, the NiCOlit dataset,[13] with more than one thousand reactions with detailed experimental conditions. We release this realistic dataset to foster the development of ML methods adapted to chemical data found in the literature. This dataset also allows us to probe meaningful questions in regard to the applications of machine learning to yield prediction. First, we compared yield prediction performances to models trained on HTE datasets. Then, we analyzed how yield prediction performs when extrapolating on new substrates, publications or coupling partners using the data structure of the NiCOlit dataset. Eventually, we selected the most relevant prediction task and derived chemical data selection rules to build an efficient training set

from literature data.

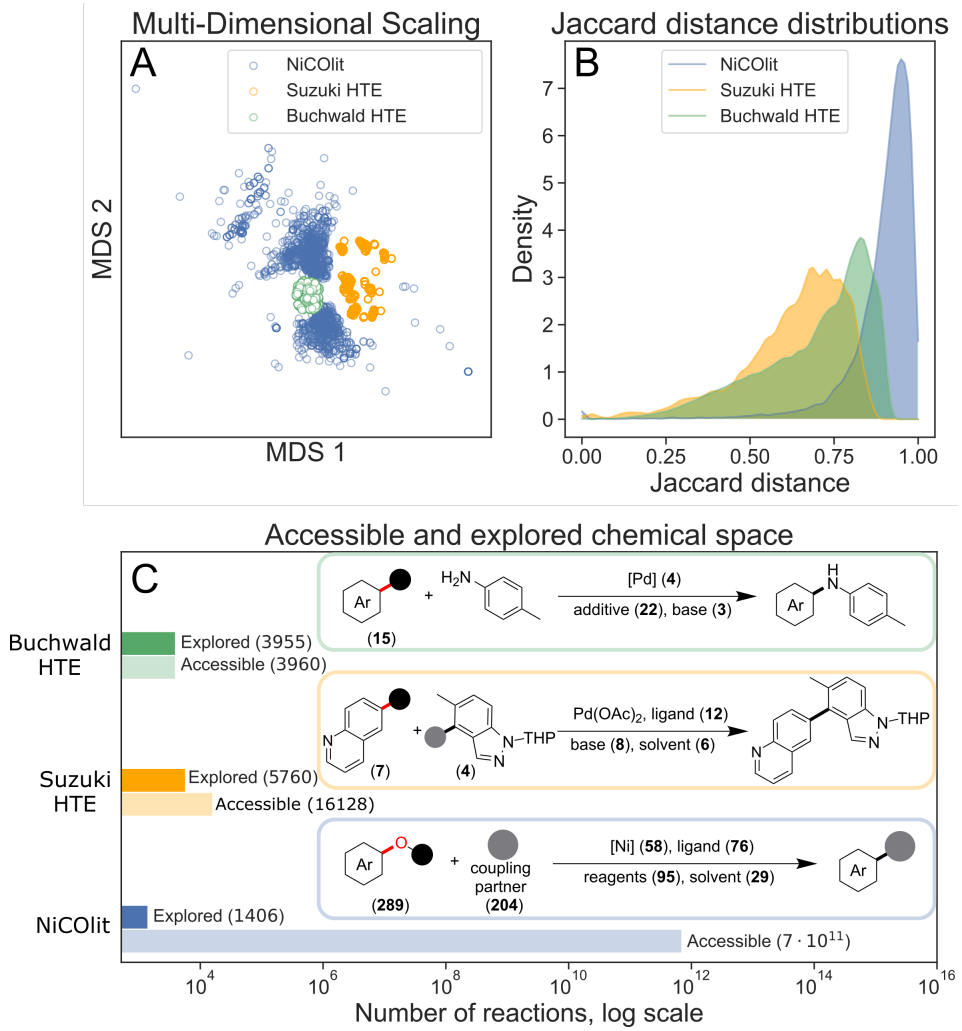


FIG. 1: Proportion of accessible chemical space observed and reaction diversities for the HTE and NiCOlit datasets. (A): Multi-Dimensional Scaling (MDS) projection of the three datasets. The NiCOlit dataset is more spread out, indicating higher chemical diversity. (B): Distribution of Jaccard distances between reactions. Distances are on average higher for the NiCOlit than the HTE datasets. Both the MDS and Jaccard distances were computed on the RXNFP representation of the reactions. (C): Proportion of accessible chemical space for the HTE and NiCOlit datasets. Numbers indicated next to each parameter corresponds to the number of different choices appearing in the dataset for this parameter. The accessible chemical space is orders of magnitude larger in the NiCOlit dataset, while the explored spaces lay in a similar range.

The NiCOlit dataset was manually extracted from literature tables and schemes cited by the review of Diao and co-workers[14]. This review focuses on the activation of carbon-oxygen bonds of phenol derivatives with nickel catalysts for coupling reactions. In order to reduce the size and the diversity of the dataset we arbitrarily restrained the study to challenging

electrophiles towards the oxidative addition: sulfonates[15] and phenols were left aside. For each reaction, the Simplified Molecular-Input Line-Entry System (SMILES)[16] chains of substrates, coupling partners, precursors, ligands, bases, additives, solvents, and products were gathered as well as experimental parameters: reaction time, temperature and molar ratio of the different partners (section II of Supplementary Materials). This highlighted current issues when harmonizing different data sources, such as disparity in yield measurement techniques, or information being reported in prose rather than machine-readable format.

The inherent differences in terms of chemical diversity and yield distributions between the NiCOlit and HTE datasets were laid out. This allows to understand how prior performances displayed by ML on the HTE data could translate to the NiCOlit data. The three datasets were projected on a common 2-dimensional space (Fig. 1A) using Multi-Dimensional Scaling[17] (see Section II of Supplementary Material). The distributions of the pairwise Jaccard distances between the reactions of each dataset (Fig. 1B) were also computed. This analysis highlights that NiCOlit reactions are much more chemically diverse than HTE reactions. Most strikingly, we calculated the accessible chemical space as the number of all possible combinations of discrete parameters used for each category (e.g. reactants, catalysts, etc.). Despite having roughly similar number of reactions in the three datasets, the accessible chemical space of NiCOlit covers almost a trillion of accessible reactions versus less than 20k for both HTE datasets (Fig. 1C). The proportion of accessible space explored[18], an intensive metric that measures the ratio between the number of chemical reactions experimentally performed and the size of the accessible chemical space, indicates a more difficult yield prediction task for the NiCOlit dataset than on HTE datasets. On the other hand, as most of the accessible space has been explored in HTE datasets (99% and 36% against only $2\text{x}10^{-7}$% for the NiCOlit dataset), developing an accurate model for the NiCOlit dataset allows to predict yields for a much larger set of unperformed reactions (almost a trillion reactions for the NiCOlit dataset).

The presence of reactions with low yields within a dataset is expected to be key for accurate predictions[11,19]. HTE datasets display relatively homogeneous yield distribution, with many negative examples. Meanwhile, searching the commercial database Sci-Finder^n (Fig. 2) for reactions matching the NiCOlit chemical reaction space returns 2,203 reactions with a clear bias toward high yields : 60% of them have a yield above 70%. The NiCOlit dataset yield distribution lays between HTEs and Sci-Finder^n data, with a significant amount of zero yields experiments but few reactions in the 20 to 40% yield range. This suggests a reporting

bias in commercial databases based on literature data.

Even though reported yield distribution for NiCOlit and HTE datasets are close, the underlying structure of the reaction data drastically differ. In HTE, all possible combinations of reactants and reaction conditions are explored (Fig. 2A). Due to time and cost constraints, chemists tend to perform a sparser exploration of the chemical space to achieve a faster convergence. Therefore, literature data is reported in two categories of tables: "optimization" and "scope". Optimization refers to the reaction conditions meaning that most parameters excepted substrate and coupling partner are modified in order to achieve an efficient reaction (vertical dots arrays Fig. 2B). In a complementary fashion, scope refers to reactions with various substrates and, or coupling partners under optimized conditions (horizontal cross arrays Fig. 2B) in order to demonstrate the robustness of the reaction. In the case of the NiCOlit dataset, we noticed that yield distribution of the optimization tables are similar to the HTE yield distribution and that scope tables display a distribution reminding that of Sci-Finder^n (Fig. 2C-D). This shows that our use of optimization tables during data extraction allows to bypass the lack of low yields reactions in literature-extracted datasets, and could afford improved predictive performances. In literature data, exploration iterates in two orthogonal directions: first, optimization of the reaction conditions, and then exploration of reactants in fixed conditions.

Then, we evaluated how existing methods for yield prediction perform on the NiCOlit dataset. As representing chemical reactions is a crucial step in statistical modeling of reaction yields[20], we selected three approaches representative of the state-of-the-art (see Section III of Supplementary Materials). The first approach, RDKit FingerPrint (RDKit FP), is based on the RDKit's (a cheminformatics tool) chemical reaction fingerprints[21,22] and one-hot encoding of the remaining variables. The second approach, referred to as Density Functional Theory (DFT), follows the guidelines given by the Auto-QChem database[23] to generate DFT-based molecular descriptors adapted to the given reaction. For the third approach, RXNFP, we featurized chemical reactions using the deep-learning RXNFP method[24]. Yield prediction models were built for each featurization with Random Forest regression models[25] that have shown state-of-the-art results on reaction yield prediction[6]. The DFT method outperforms the two others, and reaches an R^2 of 0.54, even though the difference with the RDKit FP is modest (R^2 of 0.49), while RXNFP showed the weakest performance (R^2 of 0.37) (Fig. S6). The rest of the manuscript focuses on the results obtained with the DFT model. As expected, the performance of the model trained on an optimization dataset

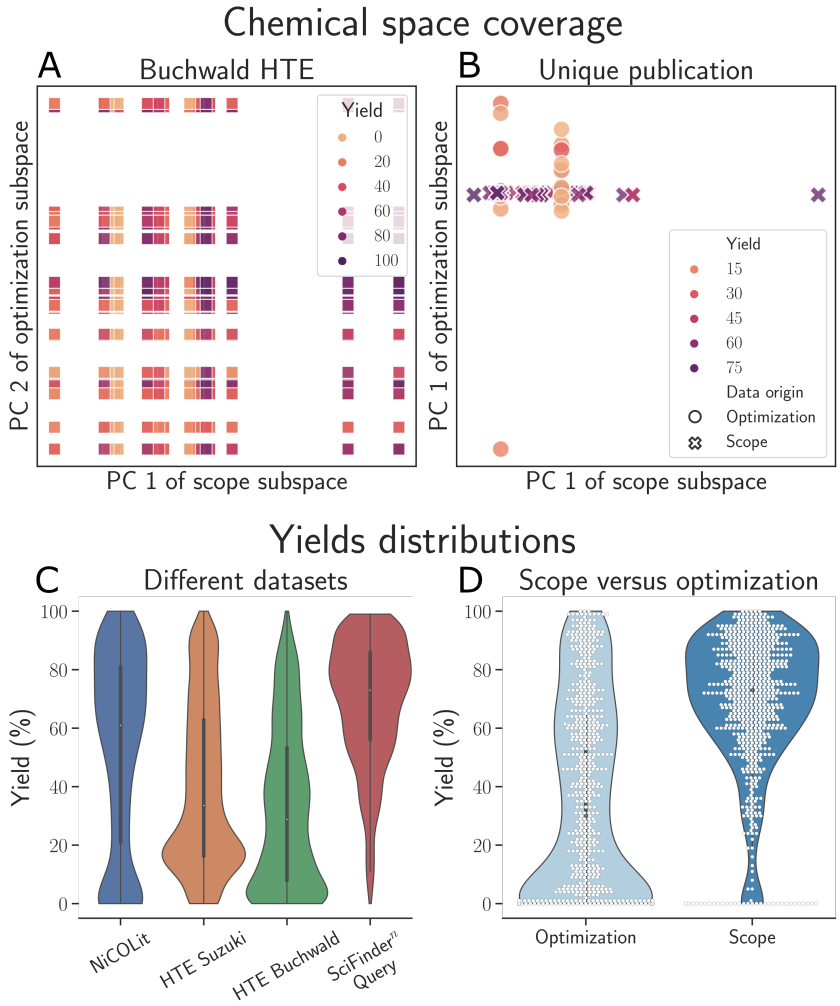


FIG. 2: Analysis of scope/optimization dataset structures and yields distributions. (**A**): Projection of the Buchwald-Hartwig dataset on the scope-optimization space, showing homogeneous coverage. The second PC is displayed as the first PC is primarily driven by the 3 bases present in the dataset. (**B**): Projection of the NiCOlit dataset on the scope-optimization space, showing a biased exploration. (**C**): Yields distribution for the NiCOlit, HTE, and Sci-Finder datasets. Bias towards high yields is observed on the NiCOlit and especially the Sci-Finder datasets. (**D**): Yields distribution for the scope and optimization data on the NiCOlit dataset.

performed better (R^2 of 0.48) than when trained on scope dataset (R^2 of 0.36) (Fig. S16).

The predictive performance on the NiCOlit dataset turns out to be far better than reported on the highly heterogeneous USPTO dataset ($R^2 < 0.2$)[7] and on data extracted from AstraZeneca's Electronic Lab Notebooks ($R^2 < 0.3$)[17]. This shows the potential of machine learning applied to reaction data extracted from the literature. Performances remain nonetheless lower than reported on the HTE datasets ($R^2 > 0.8$). An explanation

could be the highly biased structure of literature data described in Fig. 2B, while HTE datasets cover a narrow and homogeneous chemical space (Fig. 2A and 1A-B) and are devoid of experimental and reporting bias[19,26,27]. Unlike data published in the literature, HTE data systematically reports yields for all reactions including low yields, and is comprised of reactions performed in the same experimental setting. This makes them a perfect use case for statistical learning compared to the NiCOlit dataset, but with a much narrower applicability space.

Previous work on reaction yield prediction[7,10] focused mainly on predictive yields on random splits of the data. The nature of scientific discovery pushes chemists to constantly explore novel reaction chemical space. Thus, the reactions for which we want to make yield prediction are not sampled from a static distribution, but undergo continuous distribution shift[28]. For instance, chemists are often interested in reactions including a novel substrate, coupling partner or ligand (Fig. 3). Therefore, validation on a random split is not necessarily informative of how a model would perform when used by chemists in a prospective fashion. This problem was underlined in recent publications[7,10,29], where machine learning algorithms showed far worse predictive performance when applied on out-of-sample data (e.g. on reactions with a novel additive not seen in the training set). While all algorithms performed well on random splits (with an R^2 above 0.9, see Table S1), those performances dropped significantly on some out-of-samples test, with coefficient of correlations R^2 at best of 0.54 (obtained with a DFT model[6]).

We investigated whether models can be used to predict yields on a novel substrate (Fig. 3). This task seems feasible, while being of relevant practical interest. We held-out all reactions that feature a given substrate; after training on the rest of the dataset, the model predicts the yields of the held-out reactions. Those results are aggregated over all substrates in the dataset. While the DFT model showed encouraging results on the substrate split task, the question of whether the reported predictive performance ($R^2 = 0.33$) is of practical interest is not clear. Therefore, we designed a realistic classification task, where the model classifies reactions using a novel substrate in two classes, high yields (> 50) and low yields (≤ 50). This use case corresponds to the situation where a chemist wants to explore reactions with a new substrate, and relies on the model's prediction to discard low yield substrates, and to prioritize efficient ones. On this task, the DFT method reached high predictive performance (with a ROC-AUC, a performance metric for classifiers, of 0.74, see Fig. S18). This highlights a practical application of yield prediction models that can be achieved with

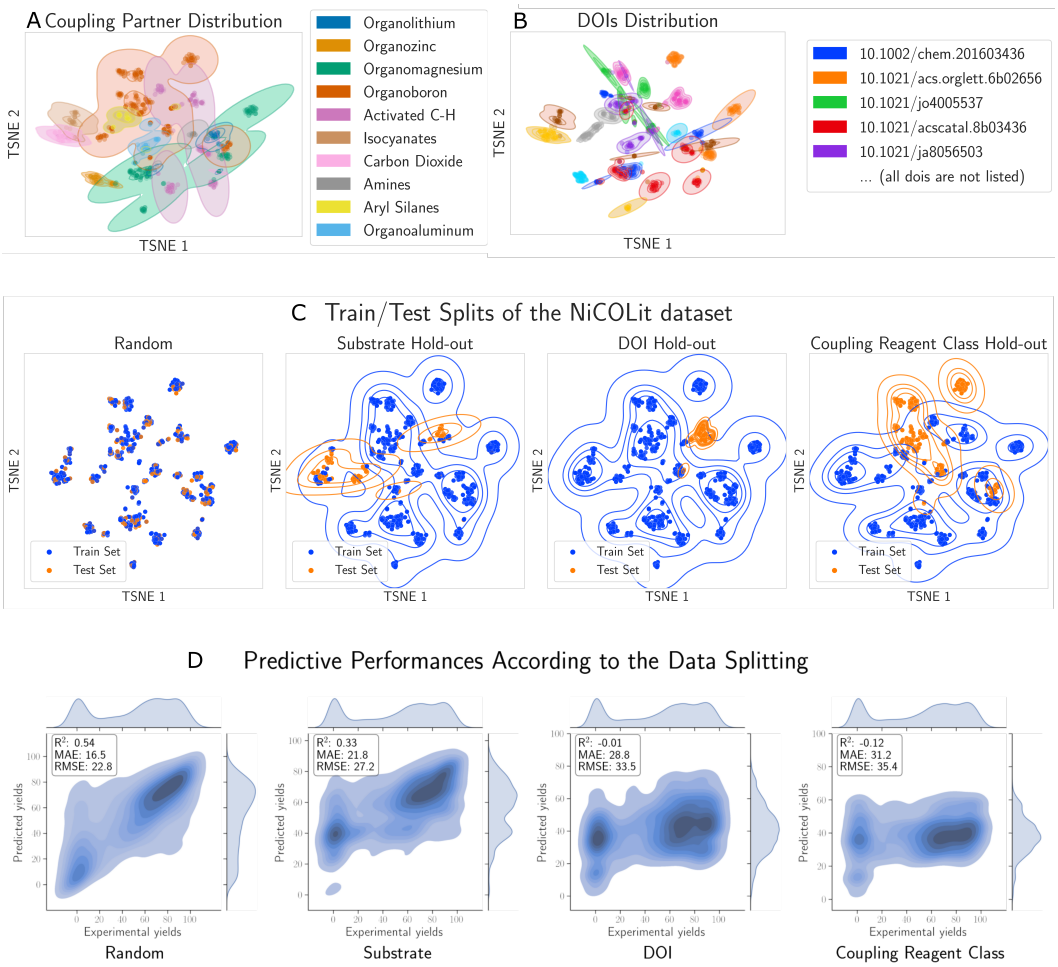


FIG. 3: ML performances on different data splits. (**A**) t-distributed stochastic neighbor embedding (t-SNE) of the NiCOlit dataset colored by coupling partner. (**B**) t-SNE of the NiCOlit dataset colored by publication. (**C**): Examples of train-test splitting from left to right : random splitting, splitting according to substrate (one substrate in the test set and all others in the training set), splitting according to publication (one Digital Object Identifier (DOI) is taken as the test set and all others as training set) and splitting according to coupling partner class (all reactions of one coupling partner class are taken as test set and all other coupling partners as training set). (**D**) DFT model performances for the splittings displayed in **C**. The performances displayed represent an average result over 10 random splits for the random task and all the possible substrate/DOI or coupling reagent class splits for the three remaining tasks.

existing methods.

Researchers' have incentives to explore novel chemical space (Fig. 3A). Moreover, for each publication, reaction yield is biased by the chemists skills and the way it is measured. This leads to a high heterogeneity between reactions from different publications. We evaluated how ML predict yields on data from a new publication. A train-test split of the data,

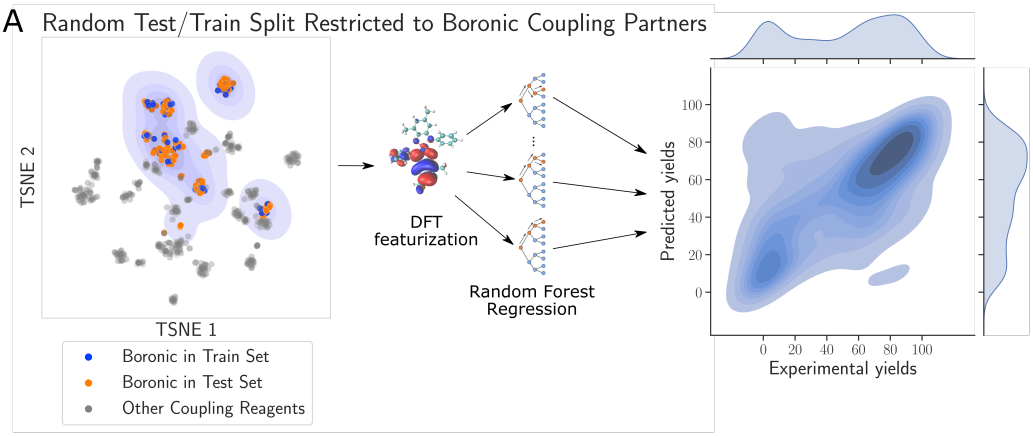
where the test set is comprised of all reactions from one publication, and the train set of all other reactions that do not appear in this publication, is used to assess the yield prediction performance of a model on a new publication (Fig. 3C - DOI Hold-out). Our results showed the inability (R^2 of -0.01) of ML to generalize to data from new publications.

While the reactions in the NiCOlit dataset are all extracted from publications referenced in the same review, they cover a wide range of possible mechanisms. As most of the publications extracted do not provide detailed mechanistic study of the reaction performed, a discrimination was made according to the nature of the coupling partner. We held-out all reactions that share a similar coupling partner (e.g. Boronic derivatives, see Fig. 3A). Models are trained on the rest of the dataset, and used to predict the yields of the held-out reactions (Fig. 3C - Coupling Partner Hold-out). Predictive performance reported for this experiment indicates whether the model is able to predict yields on coupling reactions using a different partner than the reactions of the training set. Again, the models fail to extrapolate to new coupling partners. This two experiments highlight remaining limitations for practical applications of yield prediction models.

Based on those results, we hypothesize that a restricted dataset comprised of reactions sharing similar coupling partners would lead to equivalent predictive performances than the full NiCOlit dataset. If true, this would give a precious guideline when gathering data from the literature in order to perform yield prediction.

To test this hypothesis, we trained the models on the NiCOlit dataset restricted to a given class of coupling partner (e.g. all reactions with a Boronic coupling partner), and compared the results with those obtained with a model trained on the full dataset. Indeed, for most of the coupling partner, the model trained on the restricted dataset performs as well as the model trained on the full data (table in Fig. 4).

Another observation is that the performance is highly variable according to the coupling partner. The most straightforward explanation for this behavior is the disparity between the number of reactions documented for each coupling partner class. The models exhibit poor or modest performance for coupling partners with less than 60 reactions reported. Adequate coverage of chemical space is crucial for building ML models[30]. Those results show that a dataset of a much smaller size than NiCOlit can be used to build a predictive model, provided that all reactions share the same coupling partner. They also shed light on the approximate number of reactions needed to reach satisfying predictive performance (roughly one hundred reactions). We compare the performances on the restricted sets with between



B Performances of the models on full and restricted datasets

Coupling Partner	R^2	R^2 (restricted)	Number of reactions	Number of publications	Accessible reactions
Boron derivatives	**0.47**	0.45	472	11	10^8
Activated C-Hs	**0.59**	0.56	271	3	3×10^7
Organomagnesiums	**0.51**	0.48	266	5	6×10^5
Carbon dioxide	**0.52**	0.51	87	1	7×10^3
Organozincs	0.54	**0.57**	68	2	1×10^3
Isocyanates	**0.39**	0.30	57	1	2×10^4
Organoaluminums	**0.20**	0.18	53	1	1×10^4
Arylsilanes	**0.64**	0.57	53	1	3×10^4
Organolithiums	-0.13	**0.05**	52	1	2×10^4
Amines	**0.17**	-0.05	27	1	5×10^2

FIG. 4: ML performances when trained on restricted chemical space. (**A**): Example of training a model on a subset of NiCOlit within a class of coupling partner (here Boronic derivatives). (**B**): The results are comparable to those obtained using the full dataset. This highlights that good results can be obtained in a low data regime if the reaction share a similar coupling partner.

70-500 reactions and retrieve comparable performances ($R^2 \approx 0.5$) than what was obtained on the Buchwald HTE dataset with a similar number of training points ($R^2 = 0.59$ for 150 data points)[10].

To reach its full potential, machine learning relies on high quality data. In the future, we expect that initiatives such as the Open Reaction Database[12] will provide the community with the data needed. In the meantime, there is currently a lack of public unbiased datasets of chemical reactions with detailed experimental conditions. By releasing the NiCOlit dataset, we hope to foster the development of impacting data-driven approaches for yield prediction. Our results highlight several key aspects of applying ML to predict reaction yields. A major aspect, often overlooked in recent publications, is the importance of defining tasks of real interest for chemists. Indeed, the choice of validation set in ML defines the applicability of the model evaluated. Random splitting, the most common practice for model evaluation in ML, does not give good insights on the predictive ability of the model for

tasks of practical interest. While predictive models cannot currently extrapolate to reactions from new publications or coupling partner category, we showed that the model is able to extrapolate on reactions with new substrates. This shows the practical interest of existing yield prediction models. Noteworthy, only DFT descriptors showed promising results in this context. This highlights the importance of the choice of descriptors, and paves the way for future work to improve transferability of yield prediction models on novel reactions. Our results also allowed to compare predictive performances on the widely studied HTE datasets and the data extracted from the literature. In a first analysis, performances were lower on the NiCOlit dataset. Nonetheless, with expert knowledge, high predictive performance can be attained by collecting only a limited number of reaction data. This shows that ML powered yield prediction is accessible even without access to large databases, and can have a huge impact for practitioners. This shows that by leveraging expert knowledge, predictive models can be built with small datasets including both scope and optimization tables (the latter are omitted by commercial databases). Hence, ML for yield prediction is accessible using scientific literature for data-mining, without requiring access to HTE or commercial databases.

While our results showed that more work and new approaches are needed so that ML can be applied to out-of-sample chemical reactions, they also highlight that practical applications such as predicting yields on novel substrates, or building predictive models with a reasonable amount of data, are already within reach. We hope that these findings and our open-access database will foster the adoption of machine learning for yield prediction in the chemistry community.

ACKNOWLEDGEMENTS

The authors thank Marc Bianciotto and Hervé Minoux for their thoughtful remarks and constructive feedback. **Funding:** CNRS and ENS are gratefully acknowledged for supporting J.S., L.G., R.V., M.V. The French National Association of Research and Technology (ANRT) is gratefully acknowledged for supporting M.L. (contract 2019/0821). M.L. is employed by Sanofi. **Authors contributions:** M.L. and J.S. designed the project. Y.S. and B.W. extracted the database under the supervision of J.S.. M.L., J.S., Y.S. and B.W. conducted data analysis and developed the code. M.L. and J.S. drafted the manuscript. L.G., R.V. and M.V. supervised the project and revised the manuscript. All authors read and approved the final manuscript. **Competing interests:** M.L. is a Sanofi employee

and may hold shares and/or stock options in the company. J.S., B.W., Y.S., R.V., L.G. and M.V. declares that they have no competing interests. **Data and materials availability:** All code and data used to produce the reported results can be found online at https://github.com/truejulosdu13/NiCOlit.

[1]S. Brogi, T. C. Ramalho, K. Kuca, J. L. Medina-Franco, M. Valko, *Frontiers in Chemistry* **8** (2020).

[2]V. Gallego, R. Naveiro, C. Roca, D. R. Insua, N. E. Campillo, *Molecular Diversity* (2021).

[3]L. Patel, T. Shukla, X. Huang, D. W. Ussery, S. Wang, *Molecules* **25**, 5277 (2020).

[4]C. W. Coley, W. H. Green, K. F. Jensen, *Accounts of Chemical Research* **51**, 1281 (2018).

[5]W. Beker, R. Roszak, A. Wołos, N. H. Angello, V. Rathore, M. D. Burke, B. A. Grzybowski, *Journal of the American Chemical Society* (2022).

[6]A. M. Żurański, J. I. Martinez Alvarado, B. J. Shields, A. G. Doyle, *Accounts of Chemical Research* **54**, 1856 (2021). PMID: 33788552.

[7]P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, *Machine Learning: Science and Technology* **2**, 015016 (2021).

[8]D. M. Lowe, *Ph.D. thesis, University of Cambridge* (2012).

[9]D. Perera, J. W. Tucker, S. Brahmbhatt, C. J. Helal, A. Chong, W. Farrell, P. Richardson, N. W. Sach, *Science* **359**, 429 (2018).

[10]D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **360**, 186 (2018).

[11]P. M. Pflüger, F. Glorius, *Angewandte Chemie International Edition* **59**, 18860 (2020).

[12]S. M. Kearnes, M. R. Maser, M. Wleklinski, A. Kast, A. G. Doyle, S. D. Dreher, J. M. Hawkins, K. F. Jensen, C. W. Coley, *Journal of the American Chemical Society* **143**, 18820 (2021).

[13]J. Schleinitz, M. Langevin, Y. Smail, B. Wehnert, truejulosdu13/nicolit: Nicolit dataset first release (2022).

[14]H. Diao, Z. Shi, F. Liu, *Synlett* **32**, 1494 (2021).

[15]S. Bajo, G. Laidlaw, A. R. Kennedy, S. Sproules, D. J. Nelson, *Organometallics* **36**, 1880 (2017).

[16]D. Weininger, *Handbook of Chemoinformatics* pp. 80 – 102 (2008).

[17]M. Saebi, B. Nan, J. Herr, J. Wahlers, Z. Guo, A. Zurański, T. Kogej, P.-O. Norrby, A. Doyle, O. Wiest, N. Chawla, *Chemrxiv preprint* (2021).

[18]D. Reker, E. A. Hoyt, G. J. Bernardes, T. Rodrigues, *Cell Reports Physical Science* **1**, 100247 (2020).

[19]P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, A. J. Norquist, *Nature* **533**, 73 (2016).

[20]K. Jorner, A. Tomberg, C. Bauer, C. Sköld, P.-O. Norrby, *Nature Reviews Chemistry* **5**, 240 (2021).

[21]G. Landrum, Rdkit: Open-source cheminformatics (2020).

[22]N. Schneider, D. M. Lowe, R. A. Sayle, G. A. Landrum, *Journal of Chemical Information and Modeling* **55**, 39 (2015).

[23]B. S. Andrzej Zuranski, autoqchem, https://github.com/PrincetonUniversity/auto-qchem (2020).

[24]P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, J.-L. Reymond, *Nature Machine Intelligence* **3**, 144 (2021).

[25]L. Breiman, *Machine Learning* **45**, 5 (2001).

[26]F. Strieth-Kalthoff, F. Sandfort, M. H. S. Segler, F. Glorius, *Chemical Society Reviews* **49**, 6154 (2020).

[27]W. L. Williams, L. Zeng, T. Gensch, M. S. Sigman, A. G. Doyle, E. V. Anslyn, *ACS Central Science* **7**, 1622 (2021).

[28]J. Quionero-Candela, M. Sugiyama, A. Schwaighofer, N. D. Lawrence, *Dataset Shift in Machine Learning* (The MIT Press, 2009).

[29]F. Sandfort, F. Strieth-Kalthoff, M. Kühnemund, C. Beecks, F. Glorius, *Chem* **6**, 1379 (2020).

[30]S. K. Kariofillis, S. Jiang, A. M. Żurański, S. S. Gandhi, J. I. M. Alvarado, A. G. Doyle, *Journal of the American Chemical Society* **144**, 1045 (2022).

[31]K. V. Chuang, M. J. Keiser, *Science* **362**, eaat8603 (2018).

[32]J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, G. E. Dahl, Neural message passing for quantum chemistry (2017).

[33]A. Sato, T. Miyao, K. Funatsu, *Molecular Informatics* p. 2100156 (2021).

[34]A. Yada, K. Nagata, Y. Ando, T. Matsumura, S. Ichinoseki, K. Sato, *Chemistry Letters* **47**, 284 (2018).

[35]S.-J. Kim, K. Koh, M. Lustig, S. Boyd, D. Gorinevsky, *IEEE Journal of Selected Topics in Signal Processing* **1**, 606 (2007).

[36]J. Friedman, T. Hastie, R. Tibshirani, *Journal of Statistic Software* **33**, 1 (2010).

[37]B. J. Reizman, Y.-M. Wang, S. L. Buchwald, K. F. Jensen, *Reaction Chemistry & Engineering* **1**, 658 (2016).

[38]Z. Fu, X. Li, Z. Wang, Z. Li, X. Liu, X. Wu, J. Zhao, X. Ding, X. Wan, F. Zhong, D. Wang, X. Luo, K. Chen, H. Liu, J. Wang, H. Jiang, M. Zheng, *Organic Chemistry Frontiers* **7**, 2269 (2020).

[39]K. Nakamura, M. Tobisu, N. Chatani, *Organic Letters* **17**, 6142 (2015).

[40]Z.-C. Cao, Q.-Y. Luo, Z.-J. Shi, *Organic Letters* **18**, 5978 (2016).

[41]F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Édouard Duchesnay, *Journal of Machine Learning Research* **12**, 2825 (2011).

[42]D. Merk, L. Friedrich, F. Grisoni, G. Schneider, *Molecular Informatics* **37**, 1700153 (2018).

[43]D. Bajusz, A. Rácz, K. Héberger, *Journal of Cheminformatics* **7** (2015).

[44]L. van der Maaten, G. Hinton, *Journal of Machine Learning Research* **9**, 2579 (2008). Pagination: 27.

[45]F. Furche, R. Ahlrichs, *The Journal of Chemical Physics* **117**, 7433 (2002).

[46]M. E. Casida, C. Jamorski, K. C. Casida, D. R. Salahub, *The Journal of Chemical Physics* **108**, 4439 (1998).

[47]A. D. Becke, *The Journal of Chemical Physics* **98**, 5648 (1993).

[48]C. Lee, W. Yang, R. G. Parr, *Phys. Rev. B* **37**, 785 (1988).

[49]S. H. Vosko, L. Wilk, M. Nusair, *Canadian Journal of Physics* **58**, 1200 (1980).

[50]P. J. Stephens, F. J. Devlin, C. F. Chabalowski, M. J. Frisch, *The Journal of Physical Chemistry* **98**, 11623 (1994).

[51]M. J. Frisch, *et al.*, Gaussian09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[52]M. J. Kamlet, J. L. M. Abboud, M. H. Abraham, R. W. Taft, *The Journal of Organic Chemistry* **48**, 2877 (1983).